Computing at CDF

Mark Neubauer

Massachusetts Institute of Technology

for the CDF Collaboration

- Introduction
- Computing requirements
- Central Analysis Facility
- Data Handling
- Toward the Grid
- Conclusions

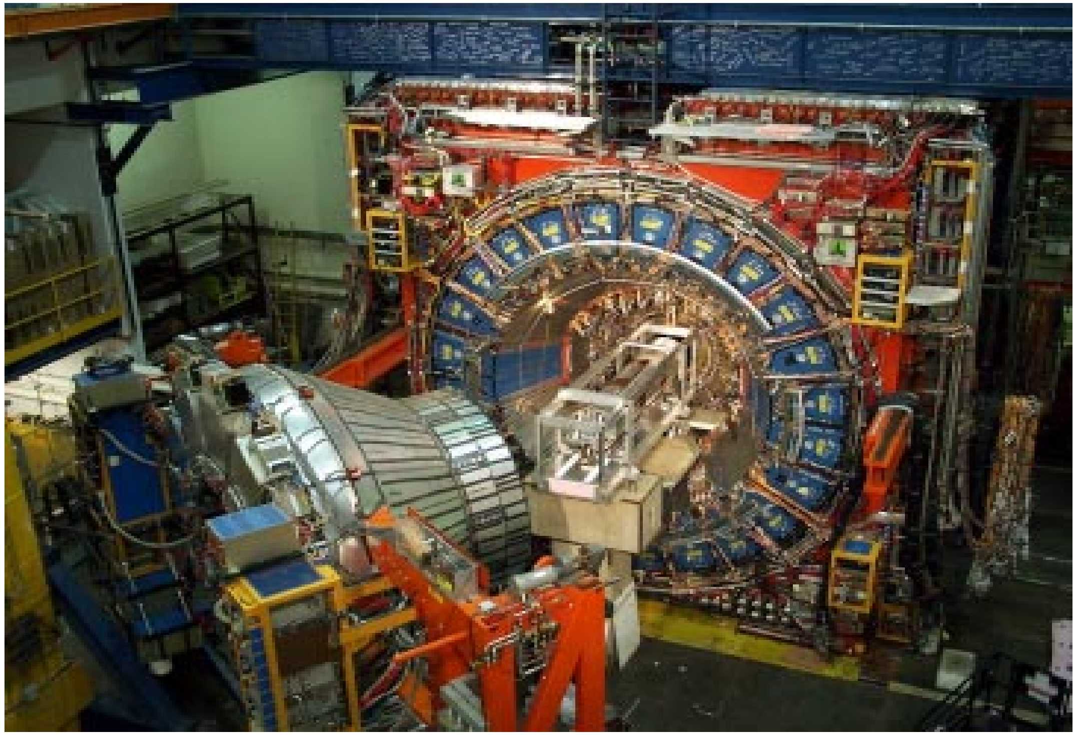
CDF in a Nutshell

- CDF + D0 experiments analyze $p\bar{p}$ collisions from Tevatron at Fermilab
- Tevatron highest energy collider in world ($\sqrt{s} = 2$ TeV) until LHC
- Run I (1992-1996) huge success → 200+ papers (t quark discovery, ...)
- Run II (March 2001-) upgrades for luminosity (×10) + energy (~10%↑)

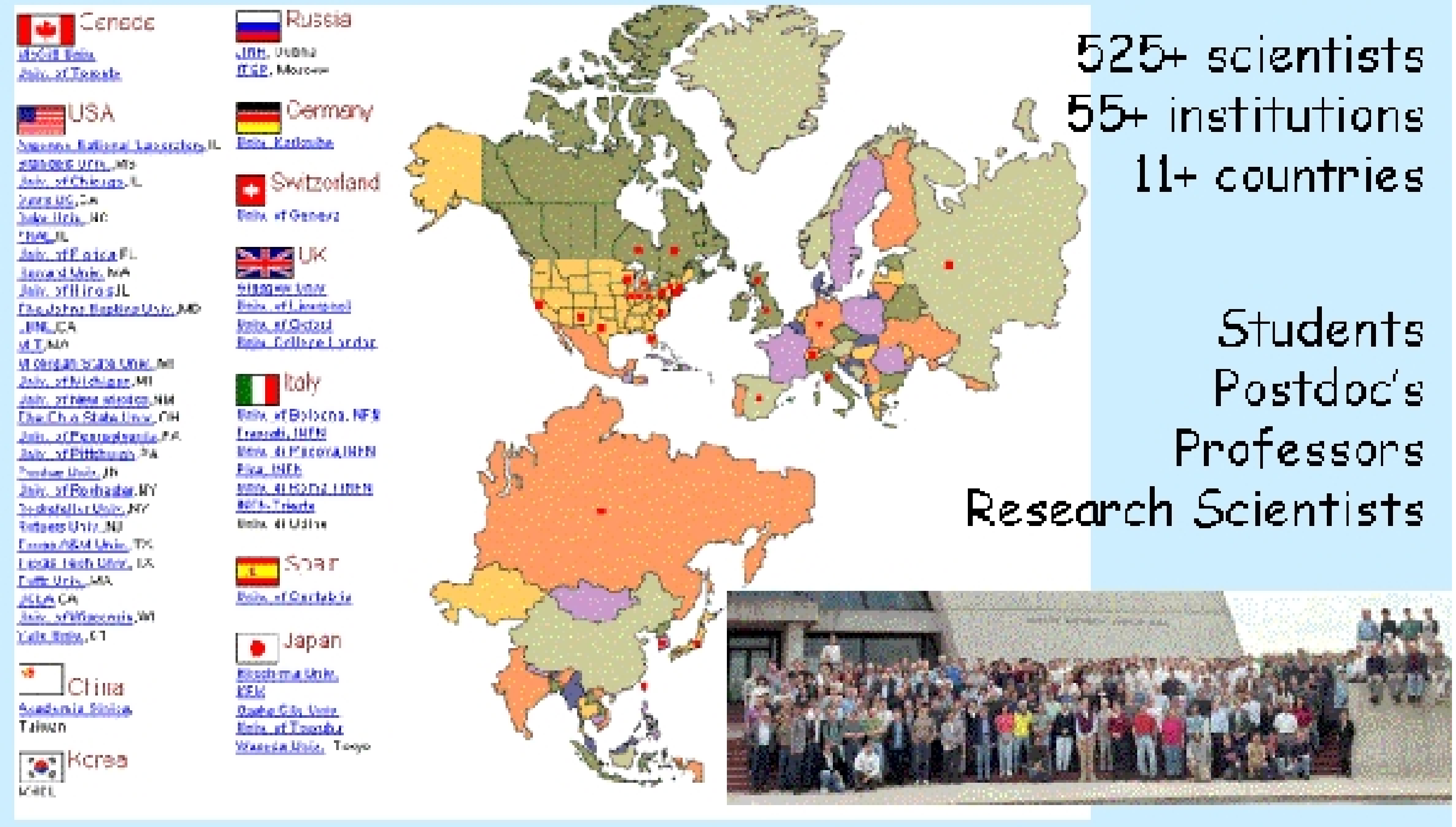
  → expect integrated luminosity 20× (Run IIa) and 150× (Run IIb) of Run I

Run II physics goals:

- **Search for Higgs boson**
- **Top quark properties** (m_t, σ_{tot}, ...)
- **Electroweak** (m_W, Γ_W, $ZZ\gamma$, ...)
- **Search for new physics** (e.g. SUSY)
- **QCD at large Q^2** (jets, α_s, ...)
- **CKM tests in *b* hadron decays**



CDF RunII Collaboration



Canada
McGill Univ.
Univ. of Toronto

USA
Argonne National Laboratory, IL
Brandeis Univ., MA
Univ. of Chicago, IL
Davis UC, CA
Duke Univ., NC
FNAL, IL
Univ. of Florida, FL
Harvard Univ., MA
Univ. of Illinois, IL
The Johns Hopkins Univ., MD
LBNL, CA
MIT, MA
Michigan State Univ., MI
Univ. of Michigan, MI
Univ. of New Mexico, NM
The Ohio State Univ., OH
Univ. of Pennsylvania, PA
Univ. of Pittsburgh, PA
Purdue Univ., IN
Univ. of Rochester, NY
Rockefeller Univ., NY
Rutgers Univ., NJ
Texas A&M Univ., TX
Texas Tech Univ., TX
Tufts Univ., MA
UCLA, CA
Univ. of Wisconsin, WI
Yale Univ., CT

China
Academia Sinica, Taiwan

Korea
KCHEP

Russia
JINR, Dubna
ITEP, Moscow

Germany
Univ. Karlsruhe

Switzerland
Univ. of Geneva

UK
Glasgow Univ.
Univ. of Liverpool
Univ. of Oxford
Univ. College London

Italy
Univ. of Bologna, INFN
Frascati, INFN
Univ. di Padova, INFN
Pisa, INFN
Univ. di Roma, INFN
INFN-Trieste
Univ. di Udine

Spain
Univ. of Cantabria

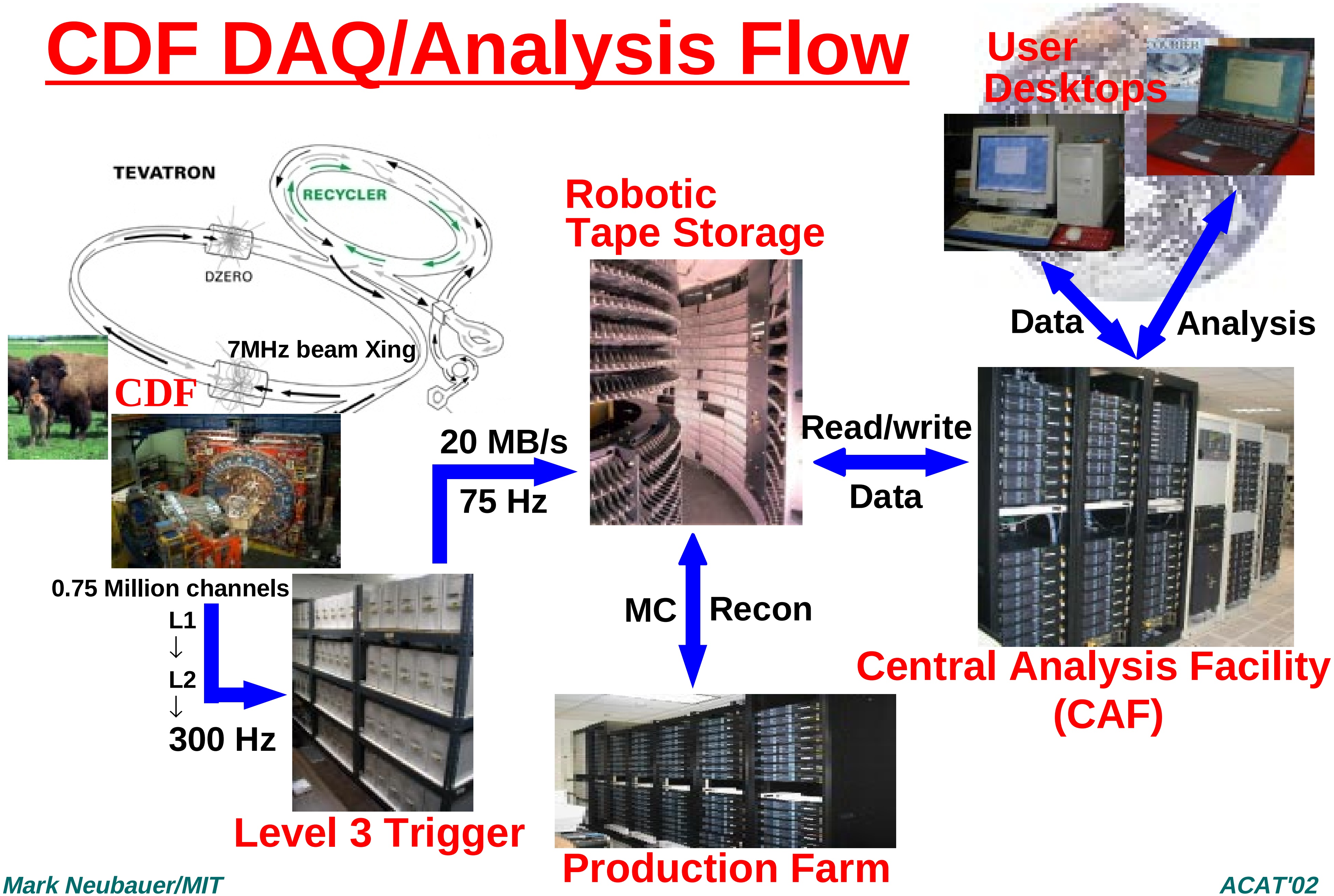
Japan
Hiroshima Univ.
KEK
Osaka City Univ.
Univ. of Tsukuba
Waseda Univ., Tokyo

525+ scientists
55+ institutions
11+ countries

Students
Postdoc's
Professors
Research Scientists

Goal: **Provide computing resources for 200+ collaborators simultaneously doing analysis per day!**

CDF DAQ/Analysis Flow



7MHz beam Xing

CDF

0.75 Million channels

L1 ↓ L2 ↓ 300 Hz

Level 3 Trigger

20 MB/s

75 Hz

Robotic Tape Storage

MC

Recon

Production Farm

Read/write

Data

Central Analysis Facility (CAF)

Data

Analysis

User Desktops

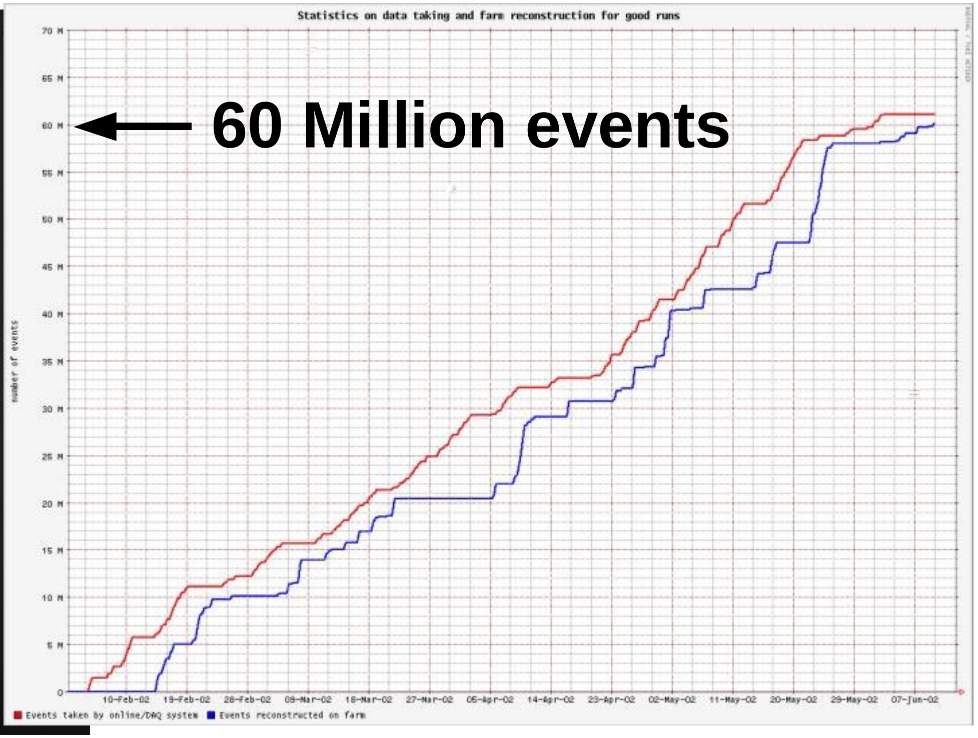
Reconstruction Farms

Data reconstruction + validation, Monte Carlo generation

154 dual P3's (equivalent to 244 1 Ghz machines)

Job management:

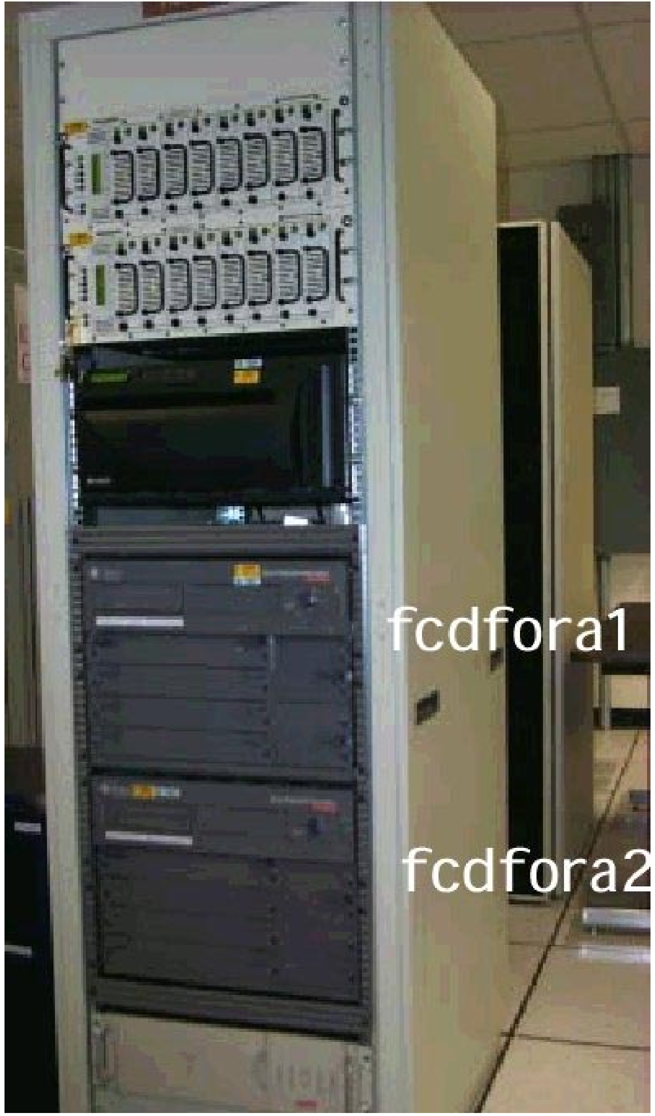
- Batch system → FBSNG developed at FNAL
- Single executable, validated offline



← 60 Million events

Database Usage at CDF

Oracle DB: Metadata + Calibrations



DB Hardware:

- **2 Sun E4500**

Presently evaluating:

- **Oracle on Linux**
- **MySQL**
- **Replication to remote sites**

Data/Software Characteristics

Data Characteristics:

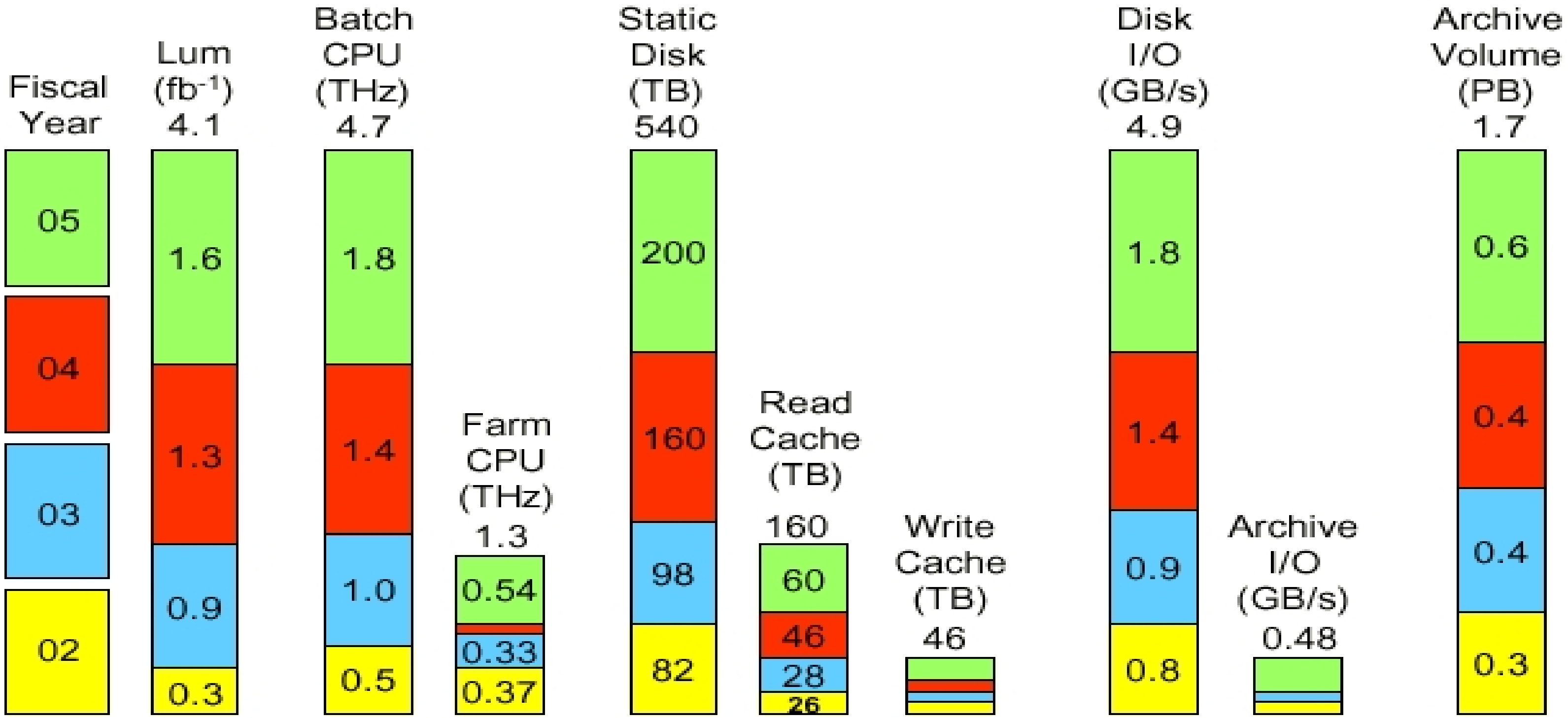
- Root I/O as persistent (raw) data format
- Raw data size: ~250 kB/event
- Reconstructed data (PAD) format: 50-100 kB/event
- Typical ntuple size (stntuple): 5-10 kB/event
- Typical RunIIa secondary dataset size: 10^7 events

Analysis Software:

- Typical analysis jobs run @ 10 Hz on 1 GHz P3

  → few MB/sec
- CPU rather than I/O bound (FastEthernet)

Computing Requirements



Requirements set by goal:

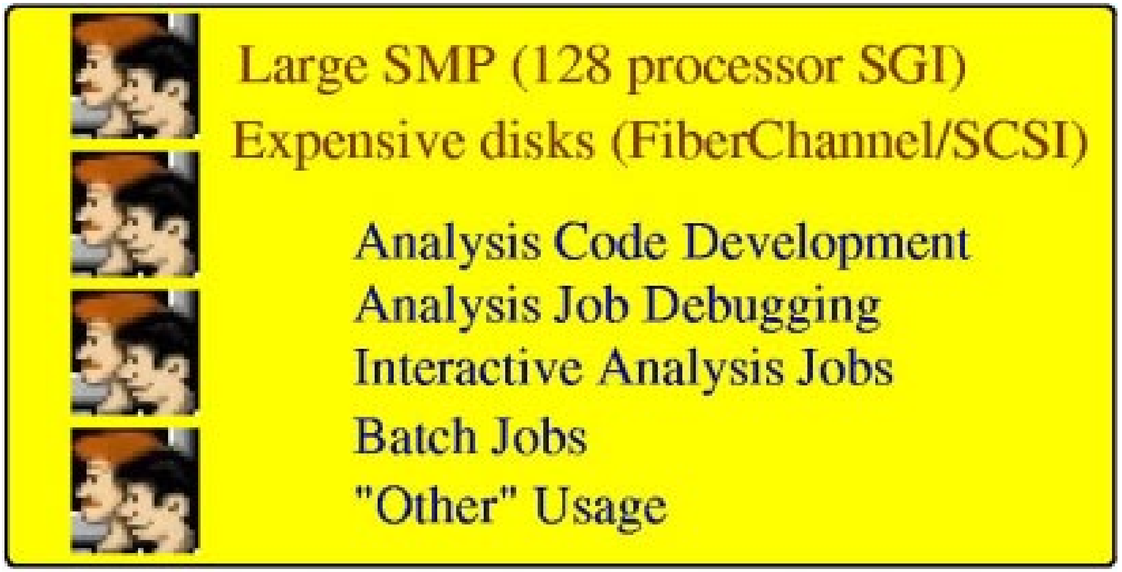
200 simultaneous users to analyze secondary data set (10^7 evts) in a day

Need ~700 TB of disk and ~5 THz of CPU by end of FY'05:

→ need lots of disk→ need cheap disk → IDE Raid

→ need lots of CPU→ commodity CPU → dual Intel/AMD

Past CAF Computing Model



Large SMP (128 processor SGI)

Expensive disks (FiberChannel/SCSI)

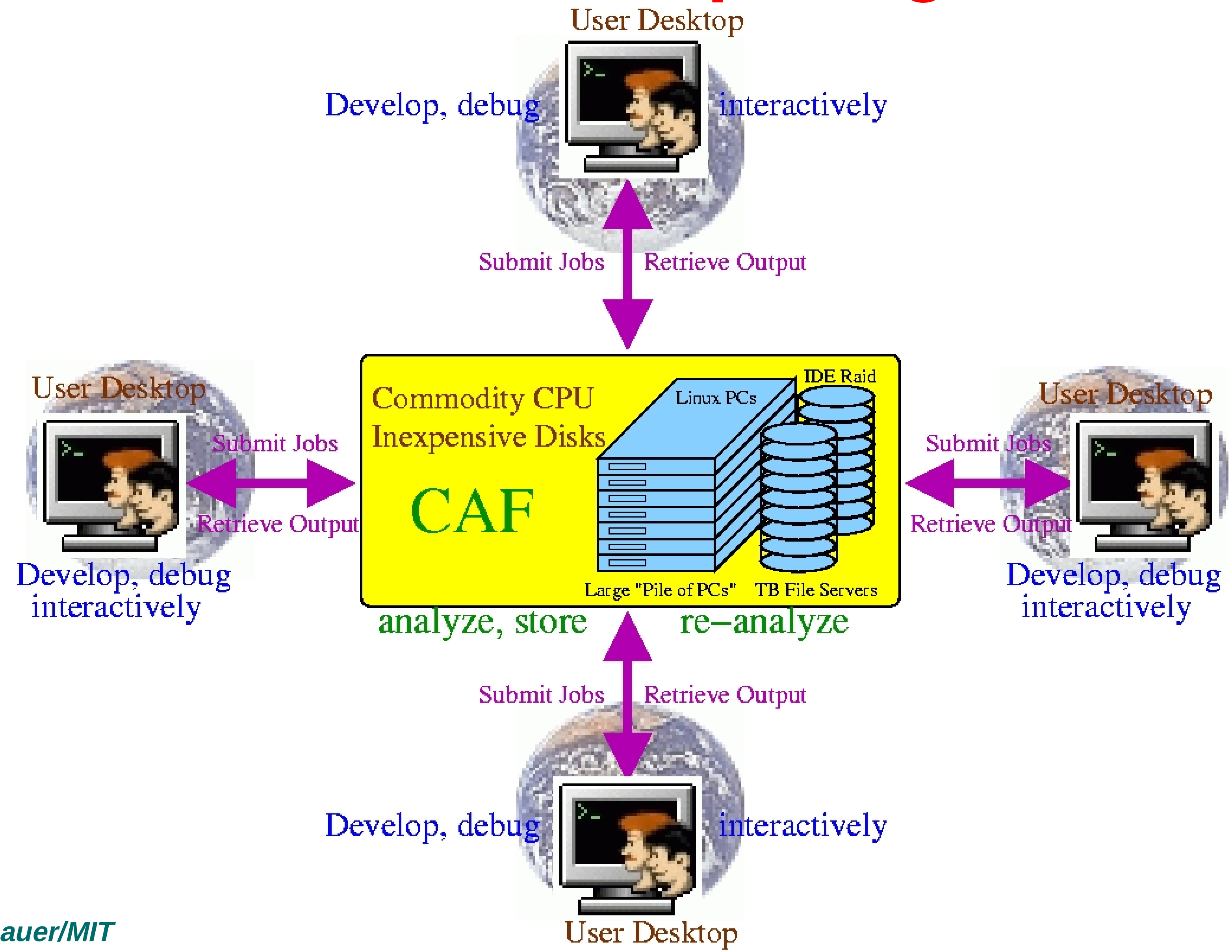
Analysis Code Development
Analysis Job Debugging
Interactive Analysis Jobs
Batch Jobs
"Other" Usage

Very expensive to expand and maintain

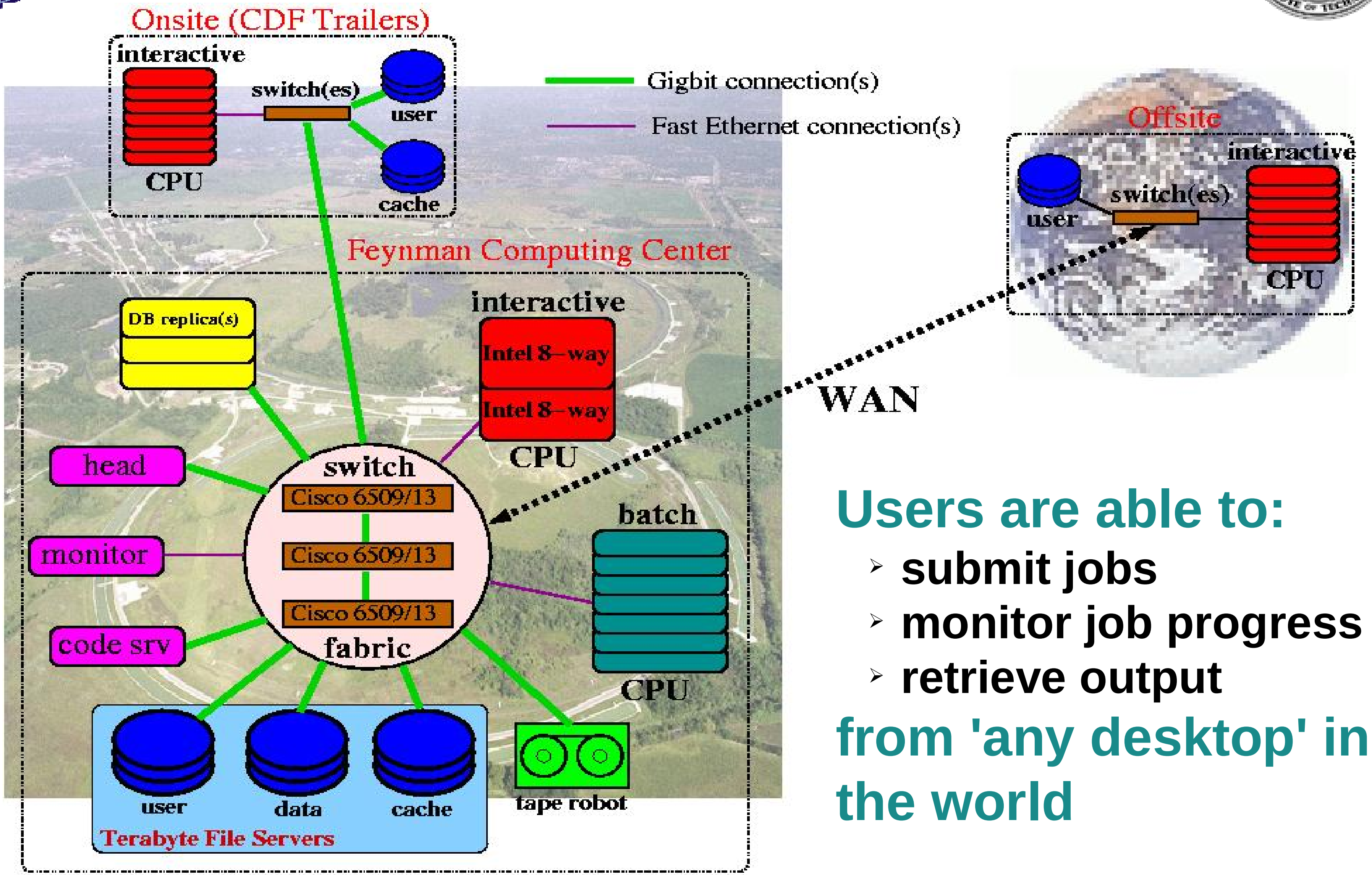
Bottom line:

Not enough 'bang for the buck'

Present CAF Computing Model



CAF Implementation



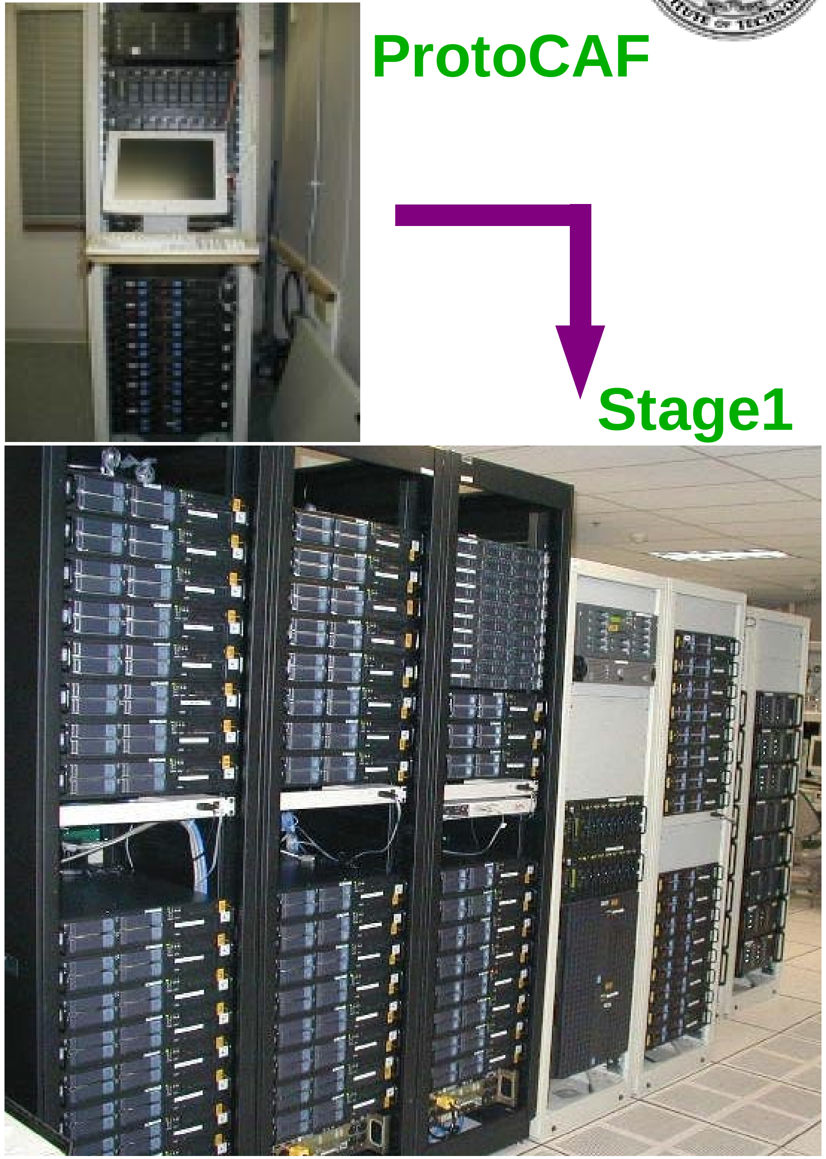
Users are able to:

- submit jobs
- monitor job progress
- retrieve output

from 'any desktop' in the world

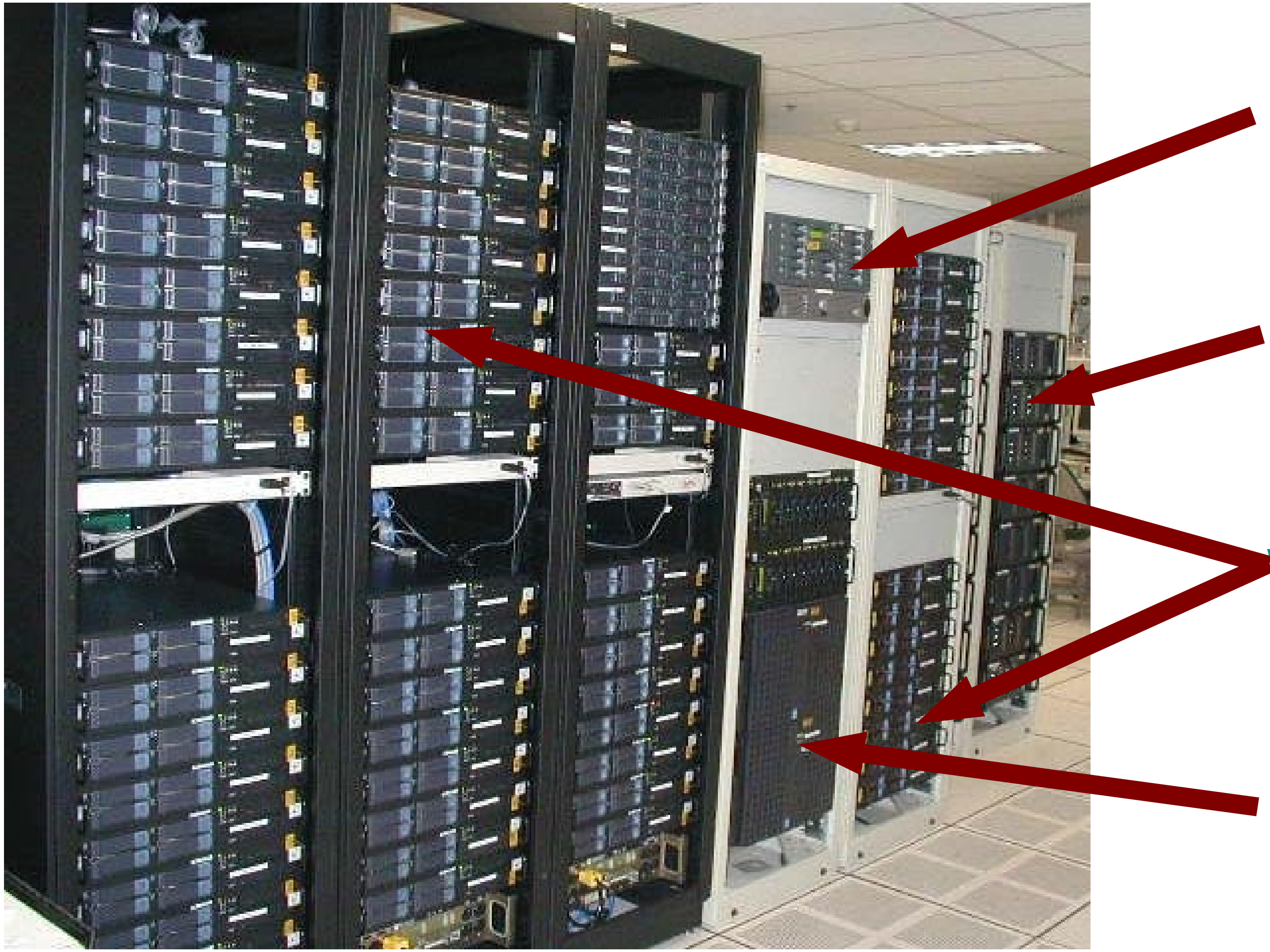
CAF Milestones

➢ Start of CAF design	11/01
➢ CAF prototype (protoCAF) assembled	2/25/02
➢ Fully-functional prototype system (>99% job success)	3/6/02
➢ ProtoCAF integrated into Stage1 system	4/25/02
➢ Production Stage1 CAF for collaboration	5/30/02



Design → Production system in 6 months!

CAF Stage 1 Hardware



Code Server

File Servers

Worker Nodes

Linux 8-ways (interactive)

Stage 1 Hardware: Workers

Workers (132 CPUs, 1U+2U rackmount):

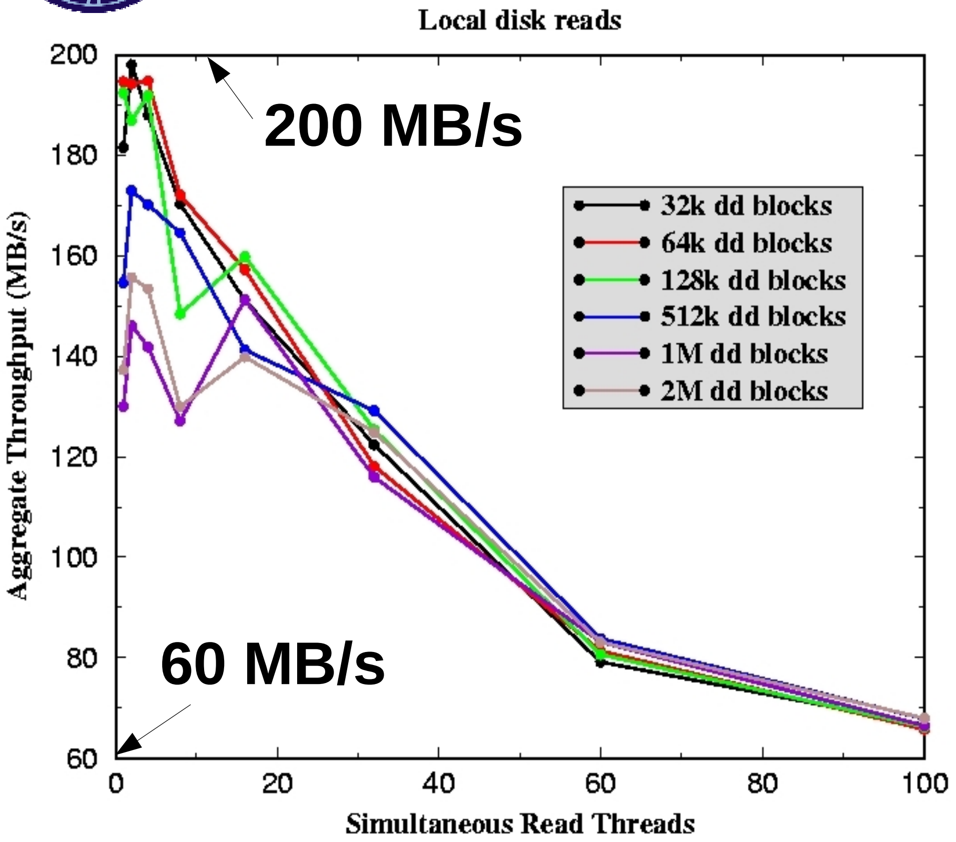
16 2U Dual Athelon 1.6GHz / 512MB RAM
50 1U/2U Dual P3 1.26GHz / 2GB RAM
FE (11 MB/s) / 80GB job scratch each

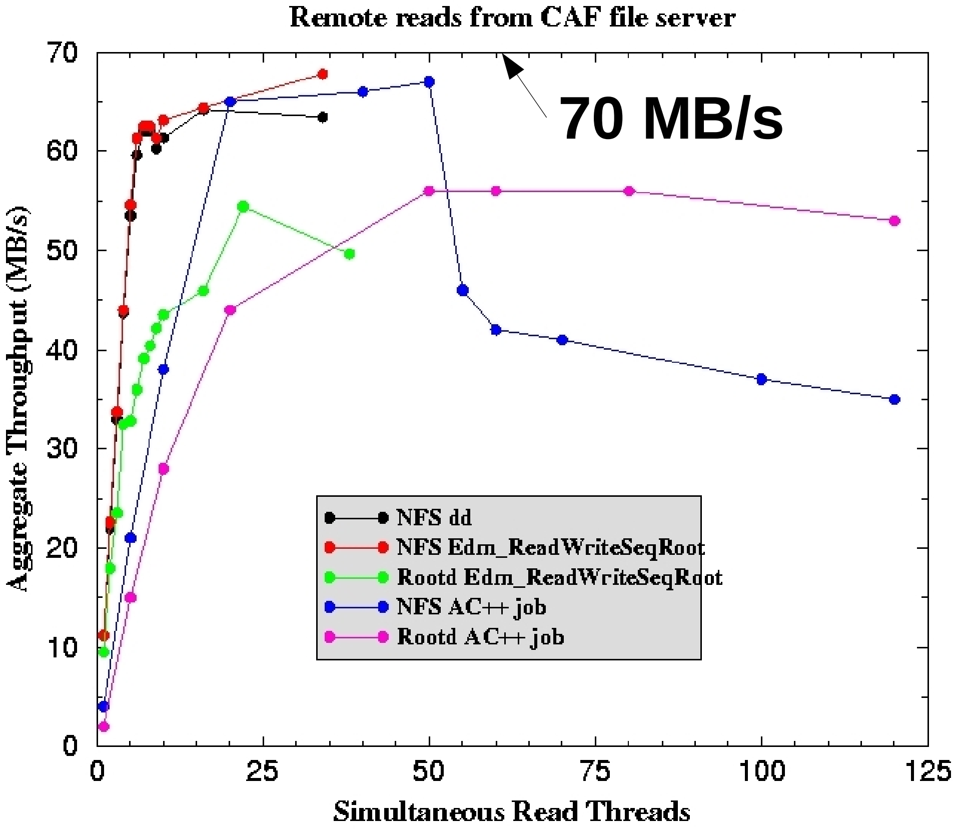
Stage 1 Hardware: Servers

Servers (35TB total, 16 4U rackmount):

- 2.2TB useable IDE RAID50 hot-swap
- Dual P3 1.4GHz / 2GB RAM
- SysKonnect 9843 Gigabt Ethernet card

File Server Performance





Server/Client Performance: Up to 200MB/s local reads, 70 MB/s NFS

Data Integrity tests: md5sum of local reads under heavy load

BER < 2×10^{-14} (Maxtor claims < 1 error / 10^{14} bits read)

Cooling tests: Temp profile of disks w/ IR gun after extended disk thrashing

CAF Software

Design goal:

Give users access to CAF resources

- CPU
- scratch disk
- data handling system

from their desktops anywhere in the world

Design constraints/desirables:

Fermilab computing security policy → kerberos!

Job scheduling → proven batch system, configurable,
fair share capability, local support → FBSNG (FNAL-CD)

Adminstrative ease → no user accounts
→ non-interactive batch, jobs run as single 'cafuser' user

User identity → unique privileges for batch jobs, disk space

User Access to CAF

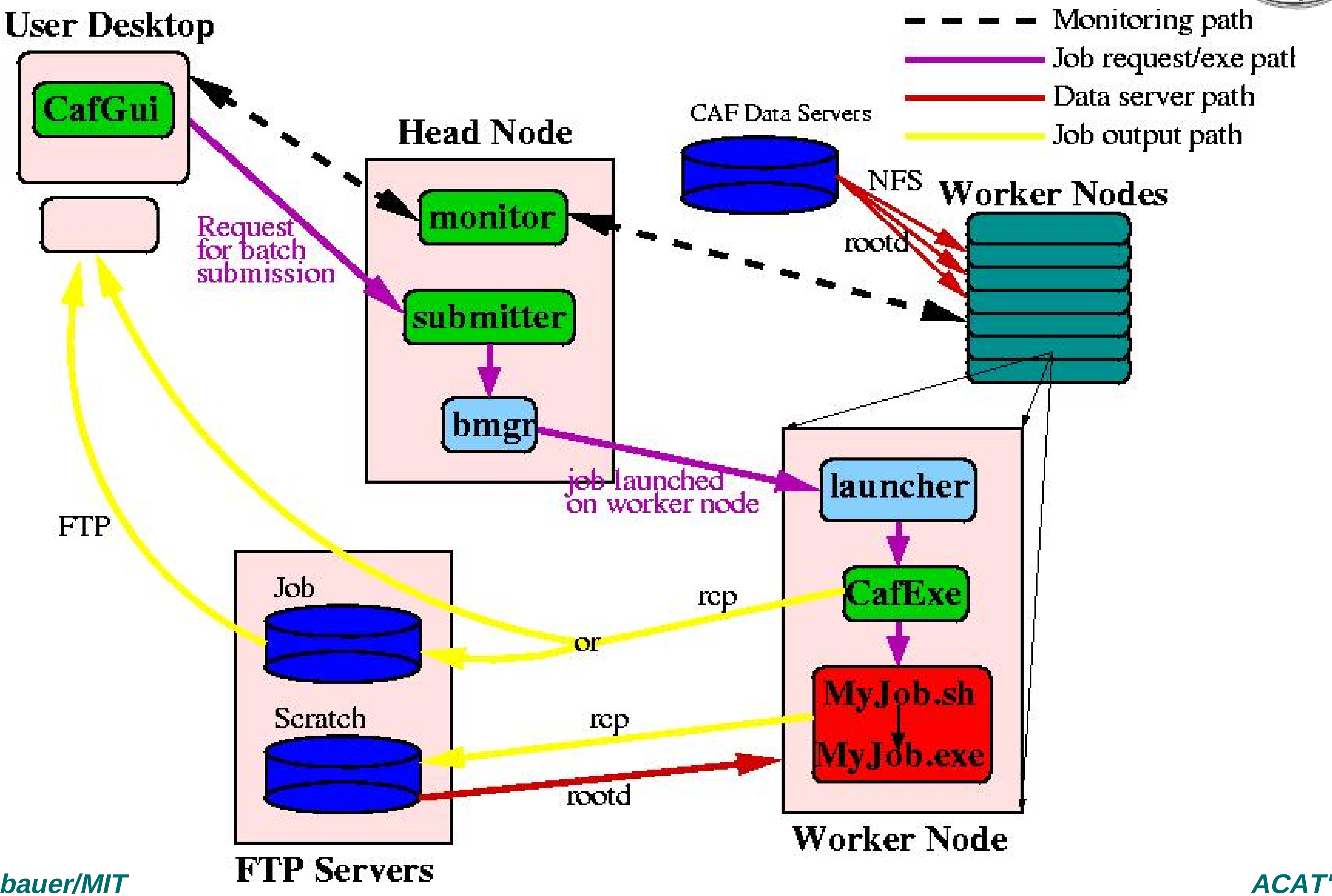
Job Related:

- Submit jobs
- Check progress of job
- Kill a job

Remote file system access:

- 'ls' in job's 'relative path'
- 'ls' in a CAF node's absolute path
- 'tail' of any file in job's 'relative path'
- Get full file listing based on metadata

CAF Software

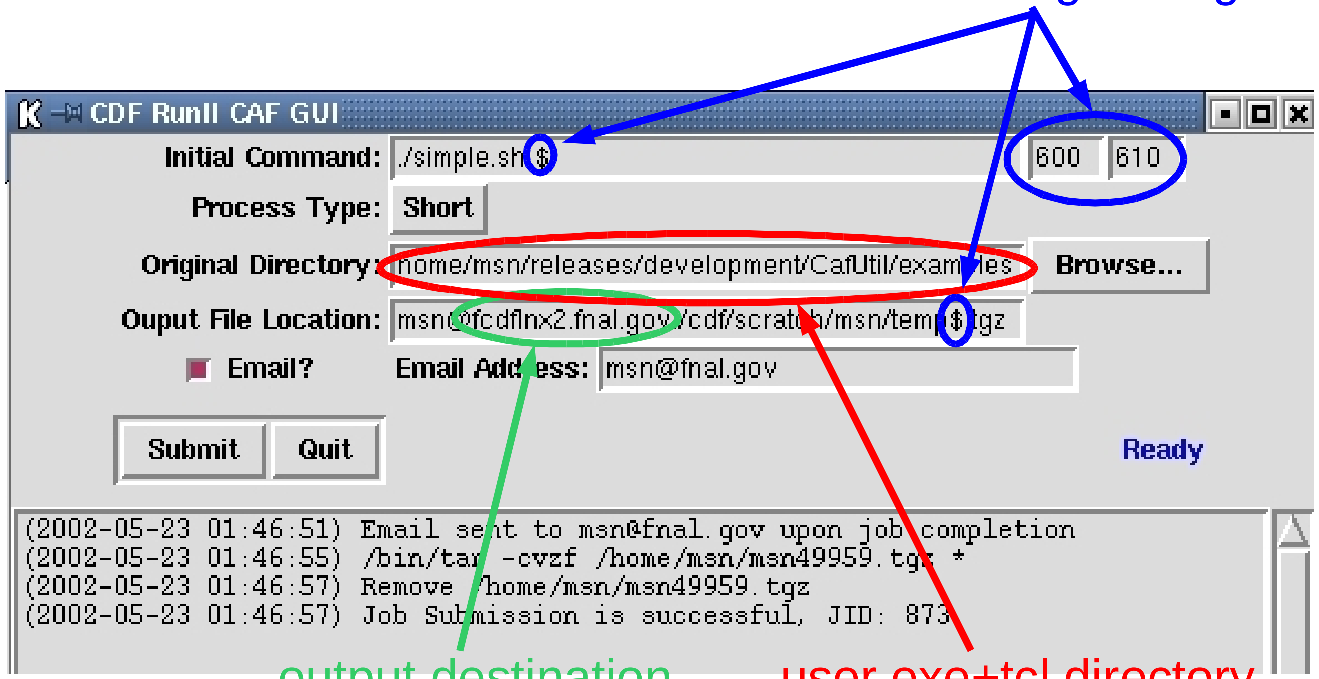


CAF User Interface

- Compile, build, debug analysis job on 'desktop'

- Fill in appropriate fields & submit job

section integer range



Initial Command: ./simple.sh $ 600 610

Process Type: Short

Original Directory: /home/msn/releases/development/CafUtil/examples

Ouput File Location: msn@fcdflnx2.fnal.gov:/cdf/scratch/msn/temp$.tgz

■ Email? Email Address: msn@fnal.gov

```
(2002-05-23 01:46:51) Email sent to msn@fnal.gov upon job completion
(2002-05-23 01:46:55) /bin/tar -cvzf /home/msn/msn49959.tgz *
(2002-05-23 01:46:57) Remove /home/msn/msn49959.tgz
(2002-05-23 01:46:57) Job Submission is successful, JID: 873
```

output destination

user exe+tcl directory

- Retrieve output using kerberized FTP tools
  ... or write output directly to 'desktop'!

Web Monitoring of User Queues

Each user a different **queue**

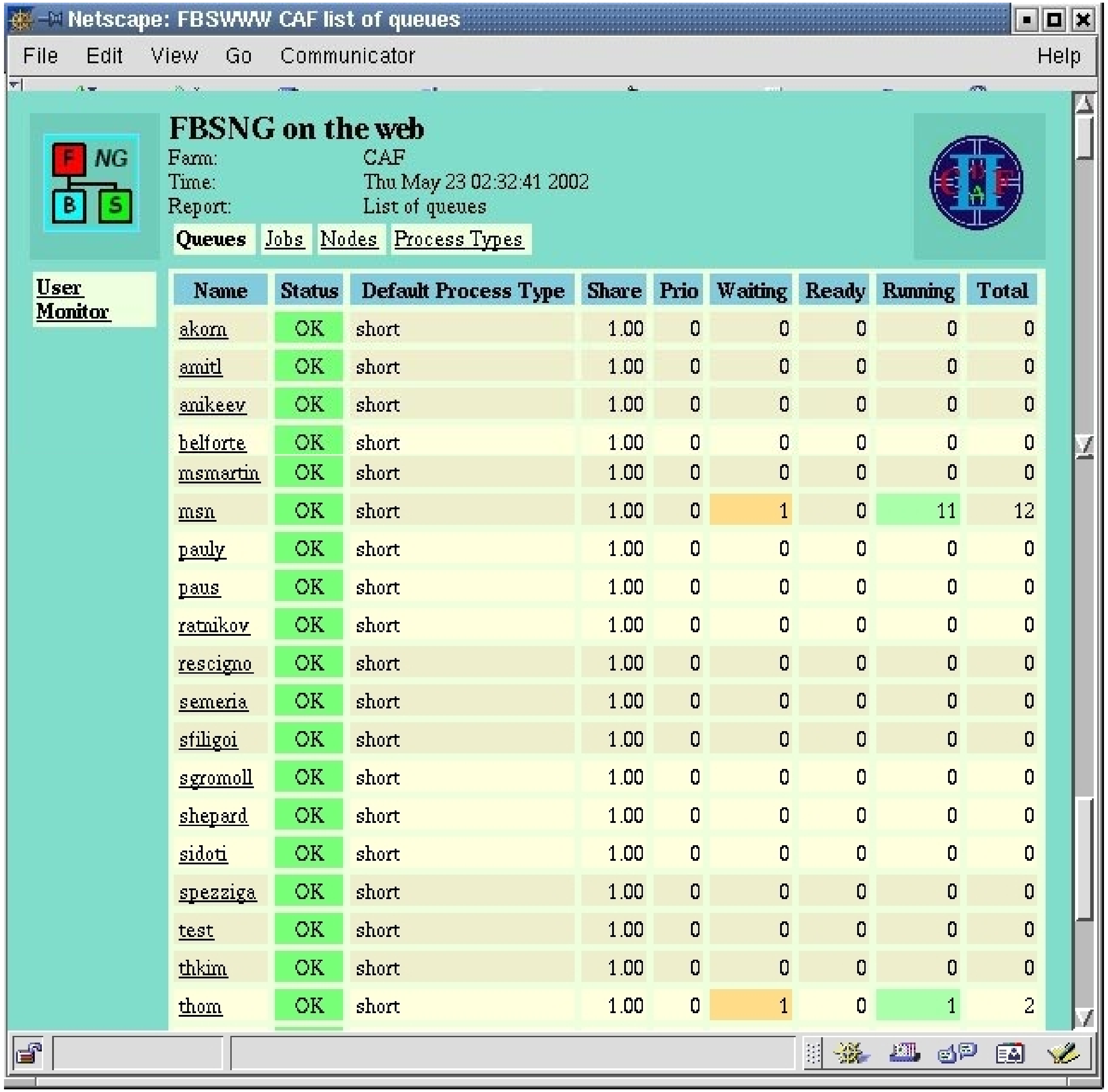
Process type for job length

- **test**: 5 mins
- **short**: 2 hrs
- **medium**: 6 hrs
- **long**: 2 days

This example:

1 job → 11 sections

(+ 1 additional section automatic for job cleanup)



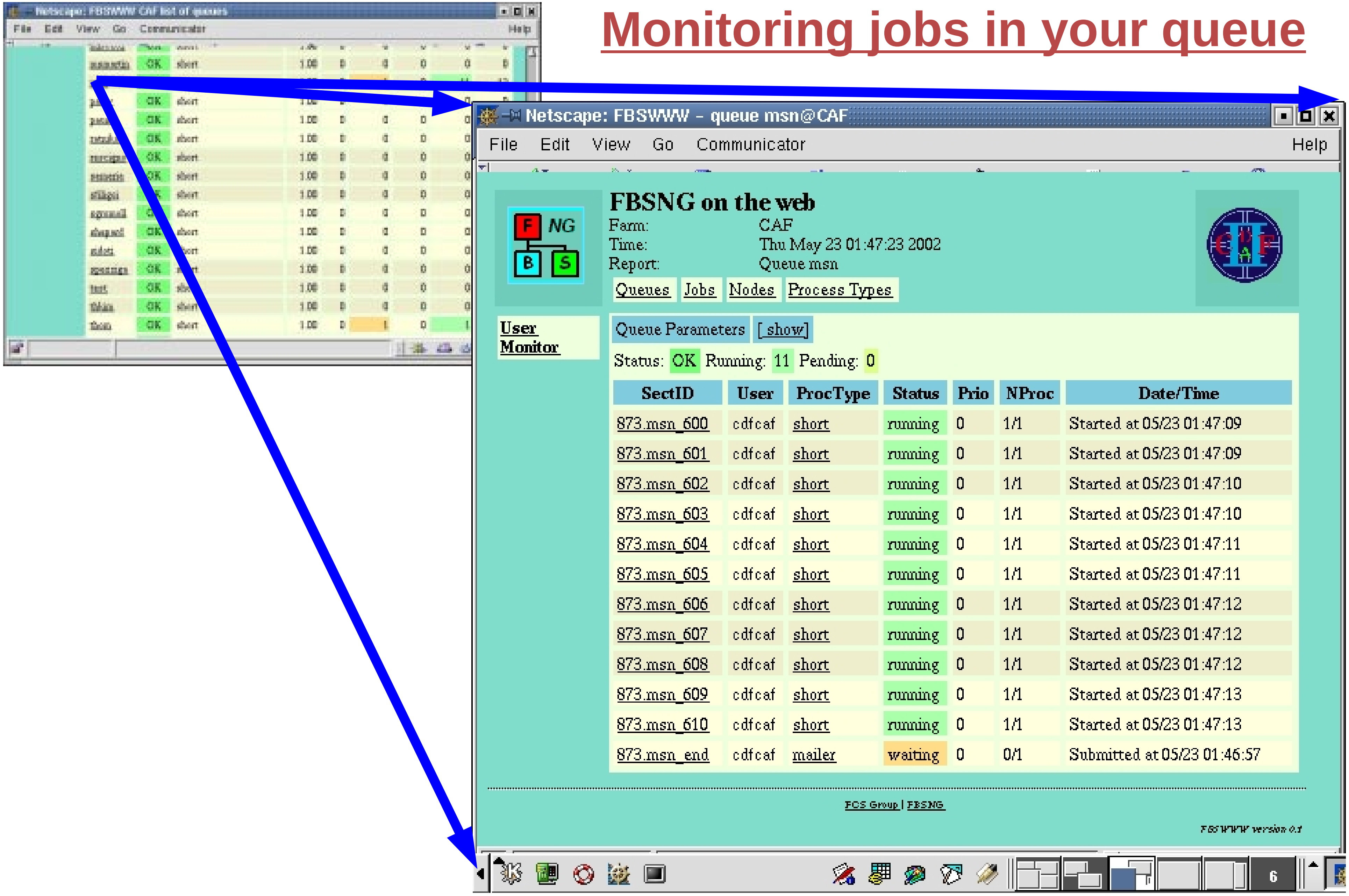
FBSNG on the web

Farm: CAF
Time: Thu May 23 02:32:41 2002
Report: List of queues

Queues | Jobs | Nodes | Process Types

Name	Status	Default Process Type	Share	Prio	Waiting	Ready	Running	Total
akorn	OK	short	1.00	0	0	0	0	0
amitl	OK	short	1.00	0	0	0	0	0
anikeev	OK	short	1.00	0	0	0	0	0
belforte	OK	short	1.00	0	0	0	0	0
msmartin	OK	short	1.00	0	0	0	0	0
msn	OK	short	1.00	0	1	0	11	12
pauly	OK	short	1.00	0	0	0	0	0
paus	OK	short	1.00	0	0	0	0	0
ratnikov	OK	short	1.00	0	0	0	0	0
rescigno	OK	short	1.00	0	0	0	0	0
semeria	OK	short	1.00	0	0	0	0	0
sfiligoi	OK	short	1.00	0	0	0	0	0
sgromoll	OK	short	1.00	0	0	0	0	0
shepard	OK	short	1.00	0	0	0	0	0
sidoti	OK	short	1.00	0	0	0	0	0
spezziga	OK	short	1.00	0	0	0	0	0
test	OK	short	1.00	0	0	0	0	0
thkim	OK	short	1.00	0	0	0	0	0
thom	OK	short	1.00	0	1	0	1	2

Monitoring jobs in your queue


Netscape: FBSWWW CAF list of queues
File Edit View Go Communicator Help
Netscape: FBSWWW – queue msn@CAF
File Edit View Go Communicator Help
F NG B S

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:47:23 2002
Report: Queue msn

Queues | Jobs | Nodes | Process Types

User Monitor

Queue Parameters [show]

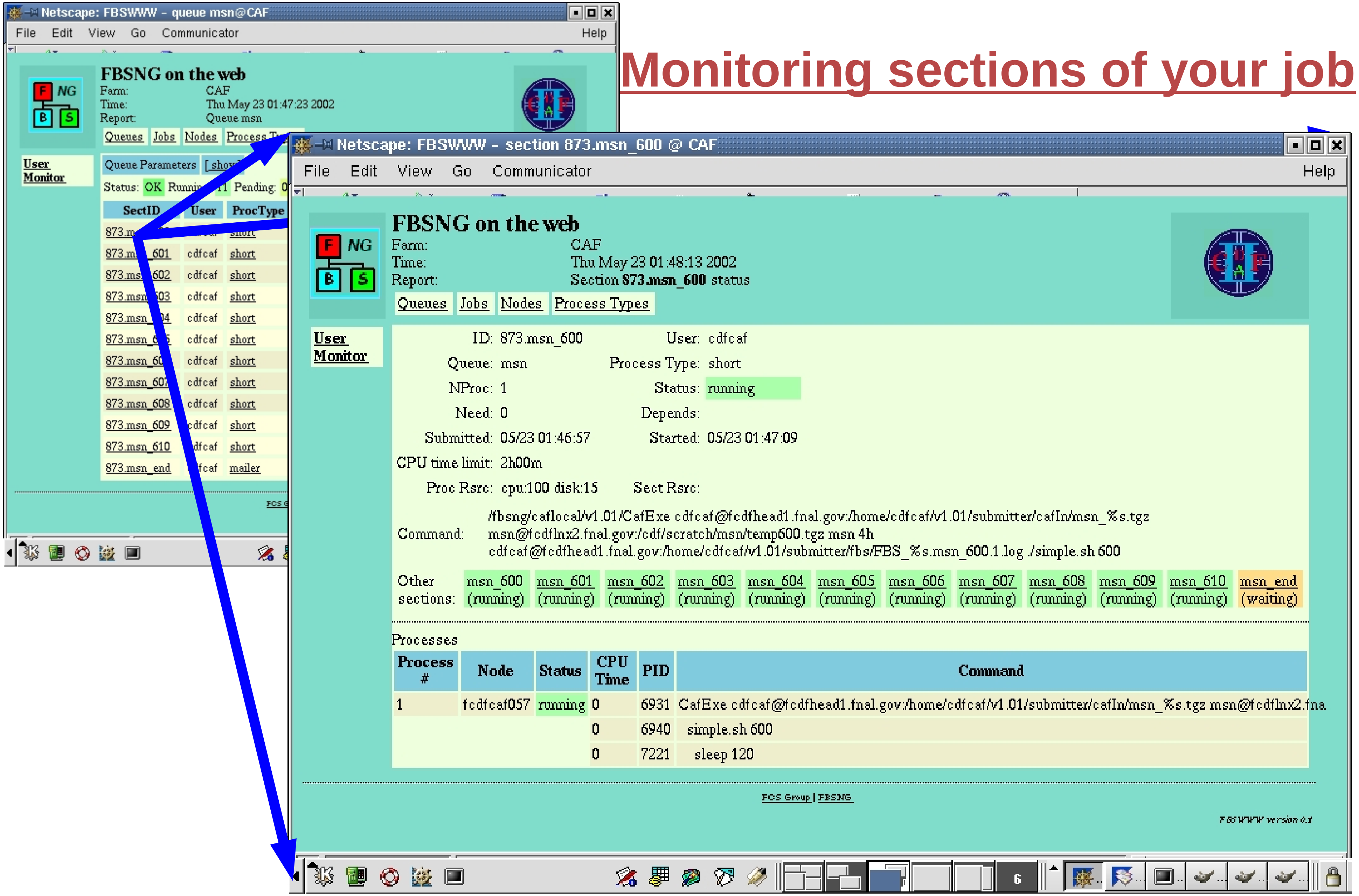
Status: OK Running: 11 Pending: 0

SectID	User	ProcType	Status	Prio	NProc	Date/Time
873.msn_600	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:09
873.msn_601	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:09
873.msn_602	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:10
873.msn_603	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:10
873.msn_604	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:11
873.msn_605	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:11
873.msn_606	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_607	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_608	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_609	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:13
873.msn_610	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:13
873.msn_end	cdfcaf	mailer	waiting	0	0/1	Submitted at 05/23 01:46:57

FCS Group | FBSNG

FBSWWW version 0.1

Monitoring sections of your job



Netscape: FBSWWW - queue msn@CAF
File Edit View Go Communicator Help

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:47:23 2002
Report: Queue msn

Queues Jobs Nodes Process Ty
User Monitor
Queue Parameters [show]

Status: OK Running: 11 Pending:

SectID	User	ProcType
873.msn_601	cdfcaf	short
873.msn_602	cdfcaf	short
873.msn_603	cdfcaf	short
873.msn_604	cdfcaf	short
873.msn_605	cdfcaf	short
873.msn_606	cdfcaf	short
873.msn_607	cdfcaf	short
873.msn_608	cdfcaf	short
873.msn_609	cdfcaf	short
873.msn_610	cdfcaf	short
873.msn_end	cdfcaf	mailer

Netscape: FBSWWW - section 873.msn_600 @ CAF
File Edit View Go Communicator Help

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:48:13 2002
Report: Section **873.msn_600** status

ID:	873.msn_600	User:	cdfcaf
Queue:	msn	Process Type:	short
NProc:	1	Status:	running
Need:	0	Depends:	
Submitted:	05/23 01:46:57	Started:	05/23 01:47:09
CPU time limit:	2h00m		
Proc Rsrc:	cpu:100 disk:15	Sect Rsrc:	

Command:
```
/fbsng/caflocal/v1.01/CafExe cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/cafIn/msn_%s.tgz
msn@fcdflnx2.fnal.gov:/cdf/scratch/msn/temp600.tgz msn 4h
cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/fbs/FBS_%s.msn_600.1.log ./simple.sh 600
```

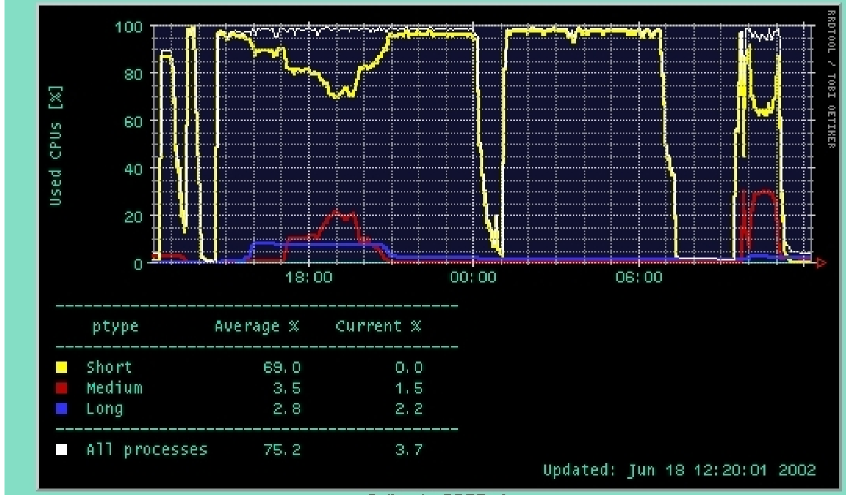
Other sections: msn_600 (running) msn_601 (running) msn_602 (running) msn_603 (running) msn_604 (running) msn_605 (running) msn_606 (running) msn_607 (running) msn_608 (running) msn_609 (running) msn_610 (running) msn_end (waiting)

Processes

Process #	Node	Status	CPU Time	PID	Command
1	fcdfcaf057	running	0	6931	CafExe cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/cafIn/msn_%s.tgz msn@fcdflnx2.fna
			0	6940	simple.sh 600
			0	7221	sleep 120

CAF Utilization

CAF global status



```
------------------------------------------------------
      ptype            Average %       Current %
------------------------------------------------------
🟨 Short                   69.0            0.0
🟥 Medium                   3.5            1.5
🟦 Long                     2.8            2.2
------------------------------------------------------
⬜ All processes           75.2            3.7
                                        Updated: Jun 18 12:20:01 2002
```

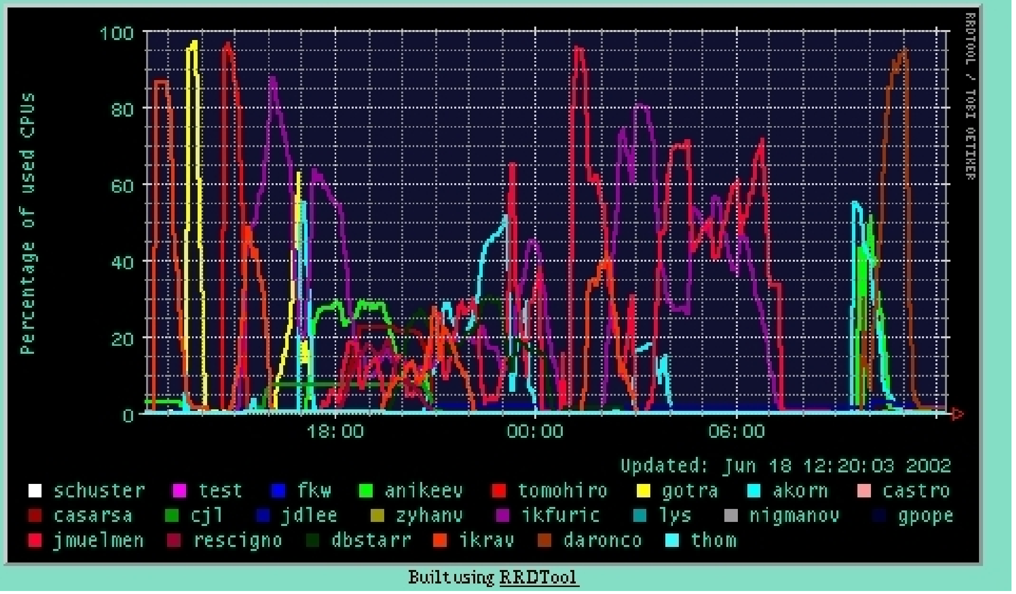
Built using RRDTool

Status summary

	Short	Medium	Long	All Types
Running sections	0	2	3	5
Pending sections	0	0	0	0
Waiting time [hh:mm] (24h average):				
per job	0:04	0:26	0:00	0:15
per section	2:12	0:52	0:00	1:32
Running time [hh:mm] (24h average)	0:20	0:35	0:00	0:27

Updated: Jun 18 12:20:02 2002

Active queues (last 24h)

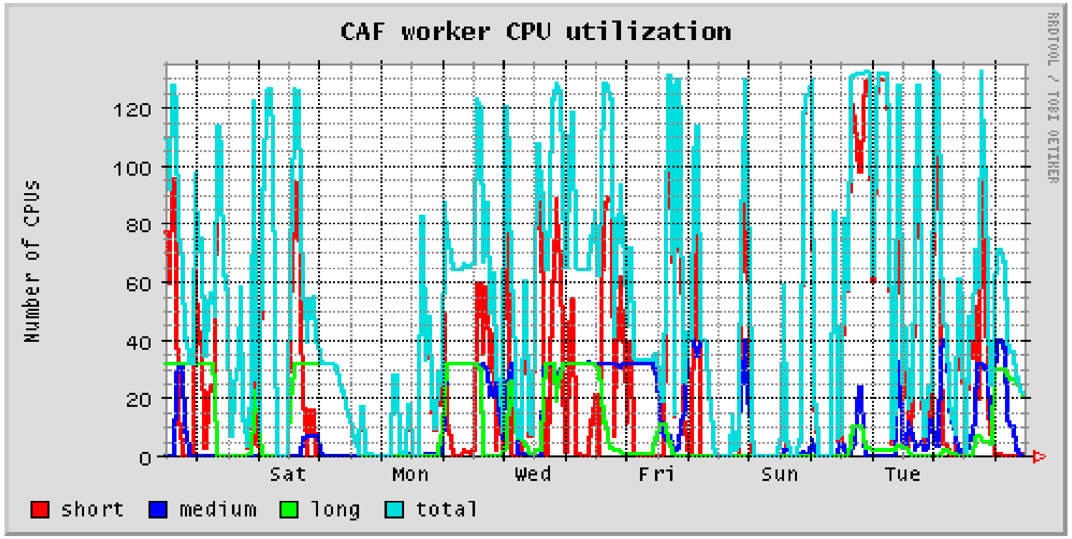


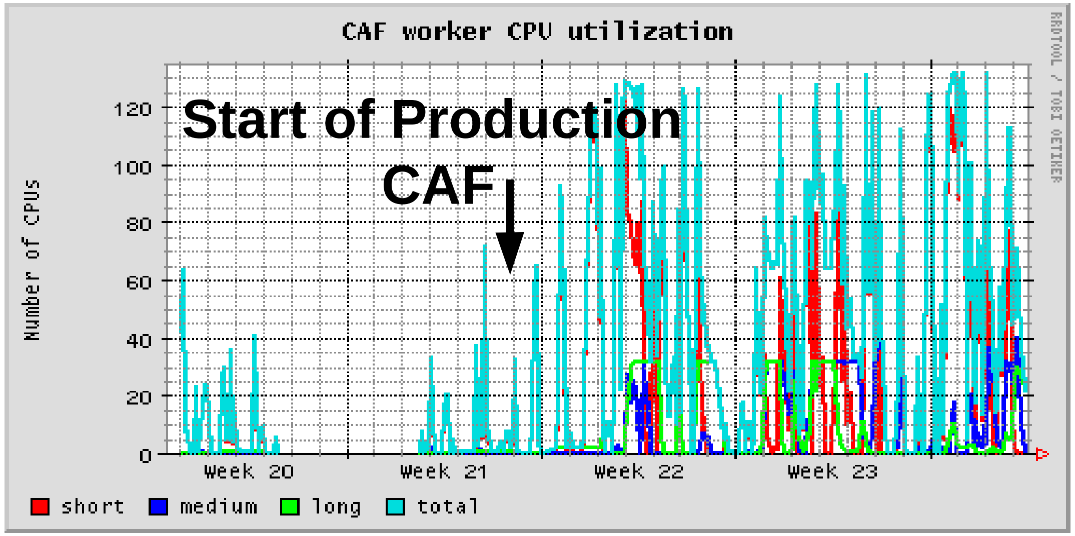
Built using RRDTool

CAF in active use by CDF collaboration

- **120 CAF Users (queues) to date**
- **2-5 new users per day**
- **Several dozen simultaneous users in a typical 24 hr period**

CAF Utilization





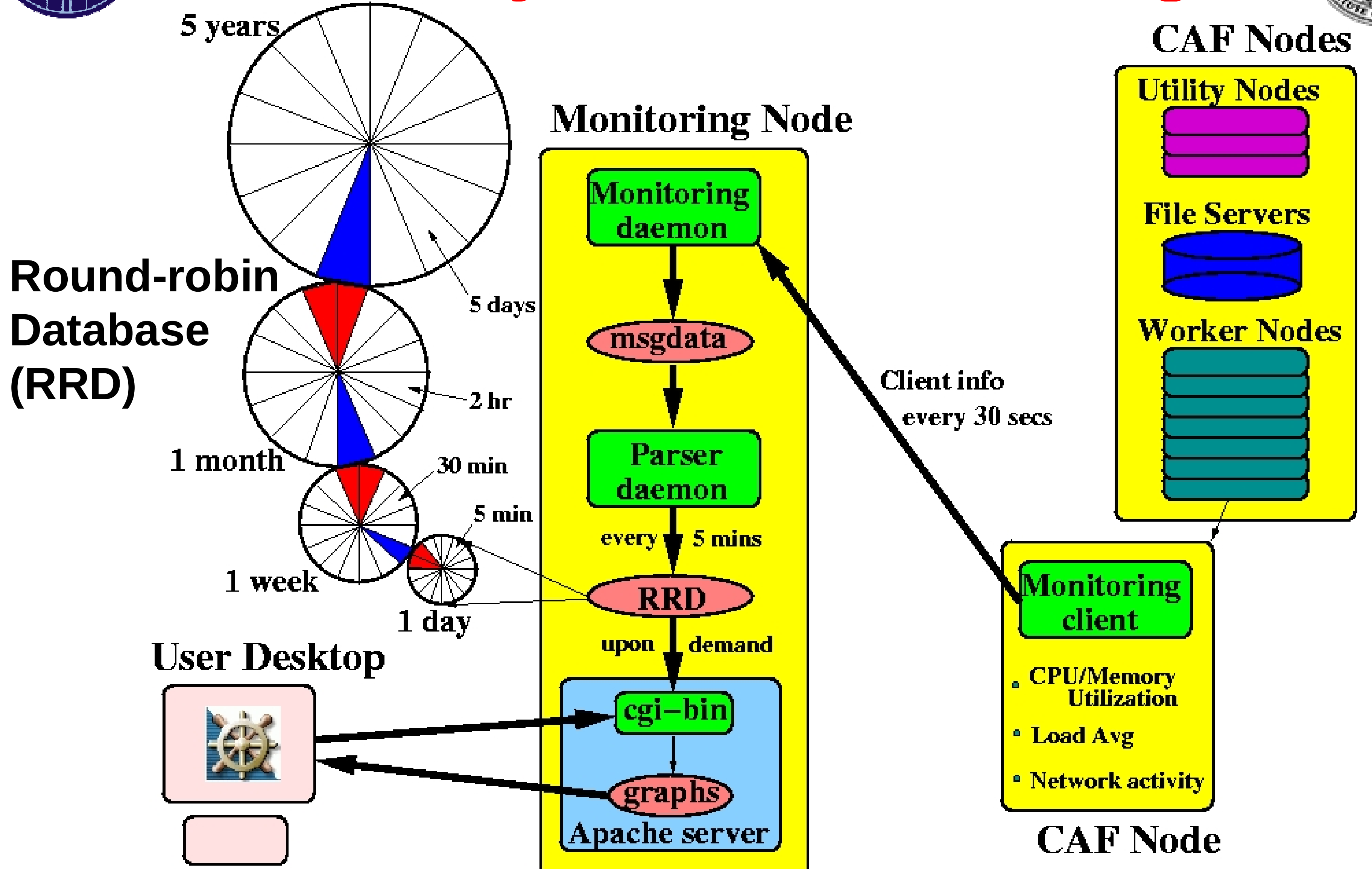
Start of Production CAF

CAF utilization steadily rising since opened to collaboration

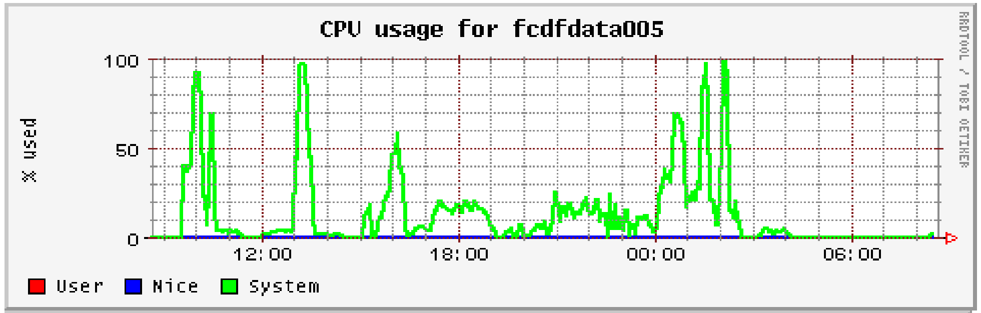
Providing 10-fold increase in analysis resources for summer physics conferences

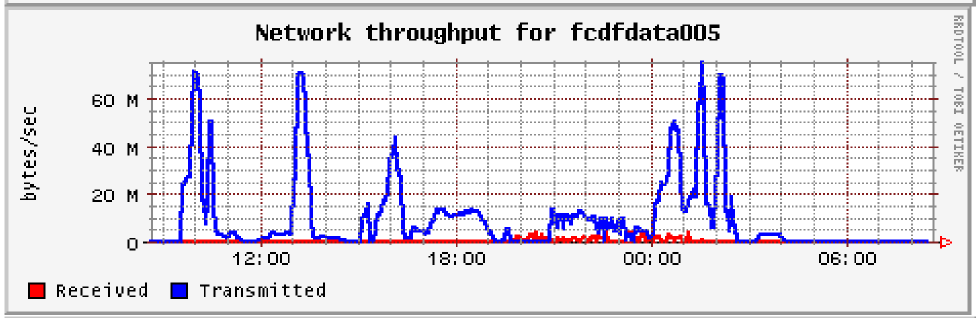
Need for more CPU on the horizon

CAF System Monitoring



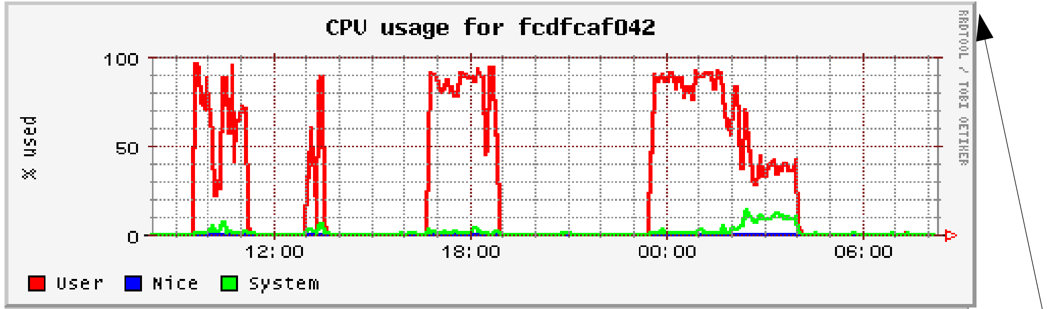
System Monitoring

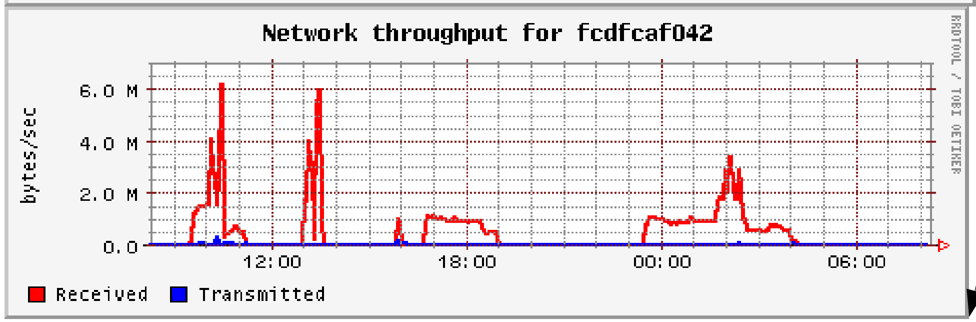




2 TB File Server

Data transfers CPU limited





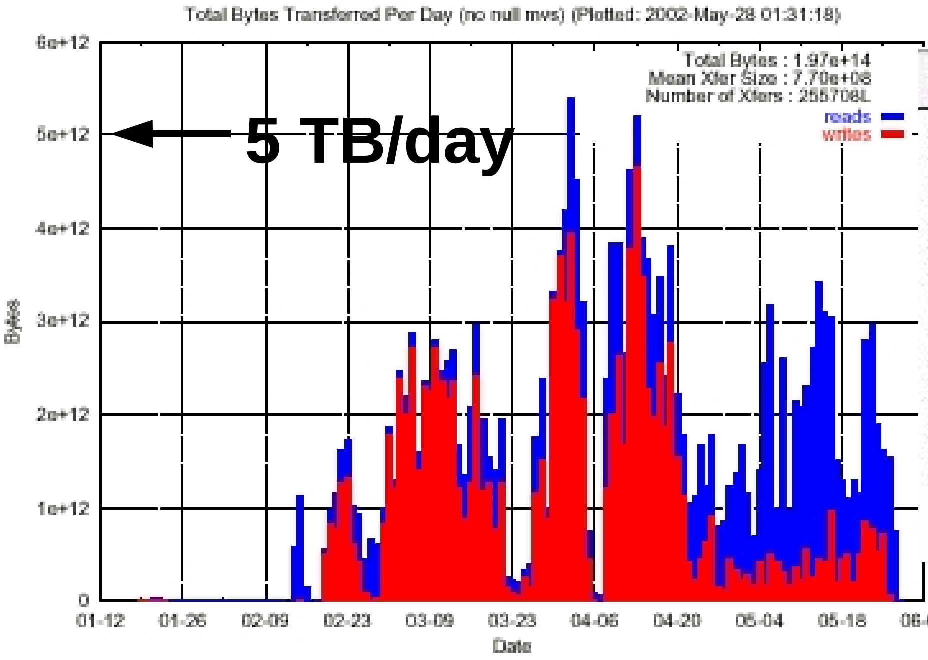
Analysis Jobs CPU bound

Worker Node

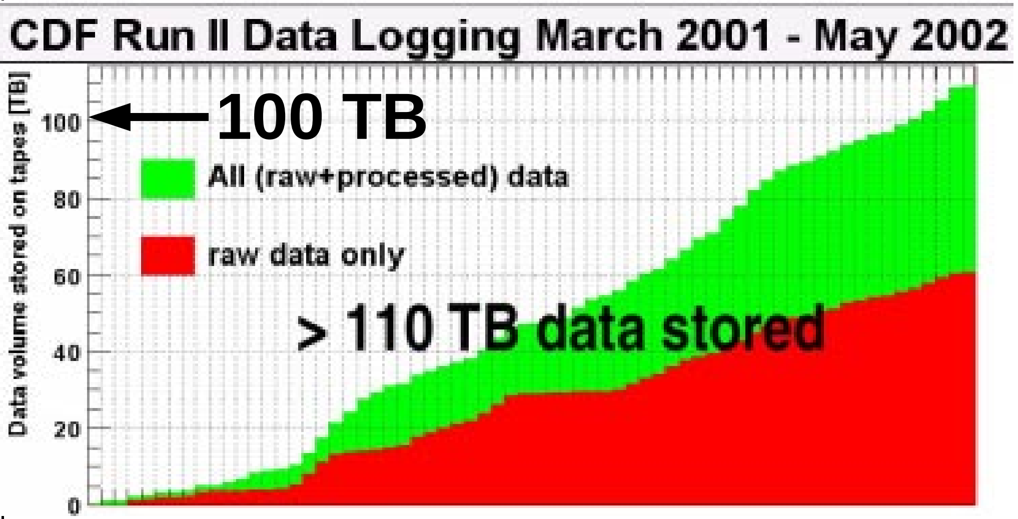
Data Handling

Data archived using STK 9940 drives and tape robot

Enstore: Network-attached tape system developed at FNAL
→ provides interface layer for staging data from tape



5 TB/day



100 TB

> 110 TB data stored

Data Handling

Dcache → network-attached disk cache from DESY

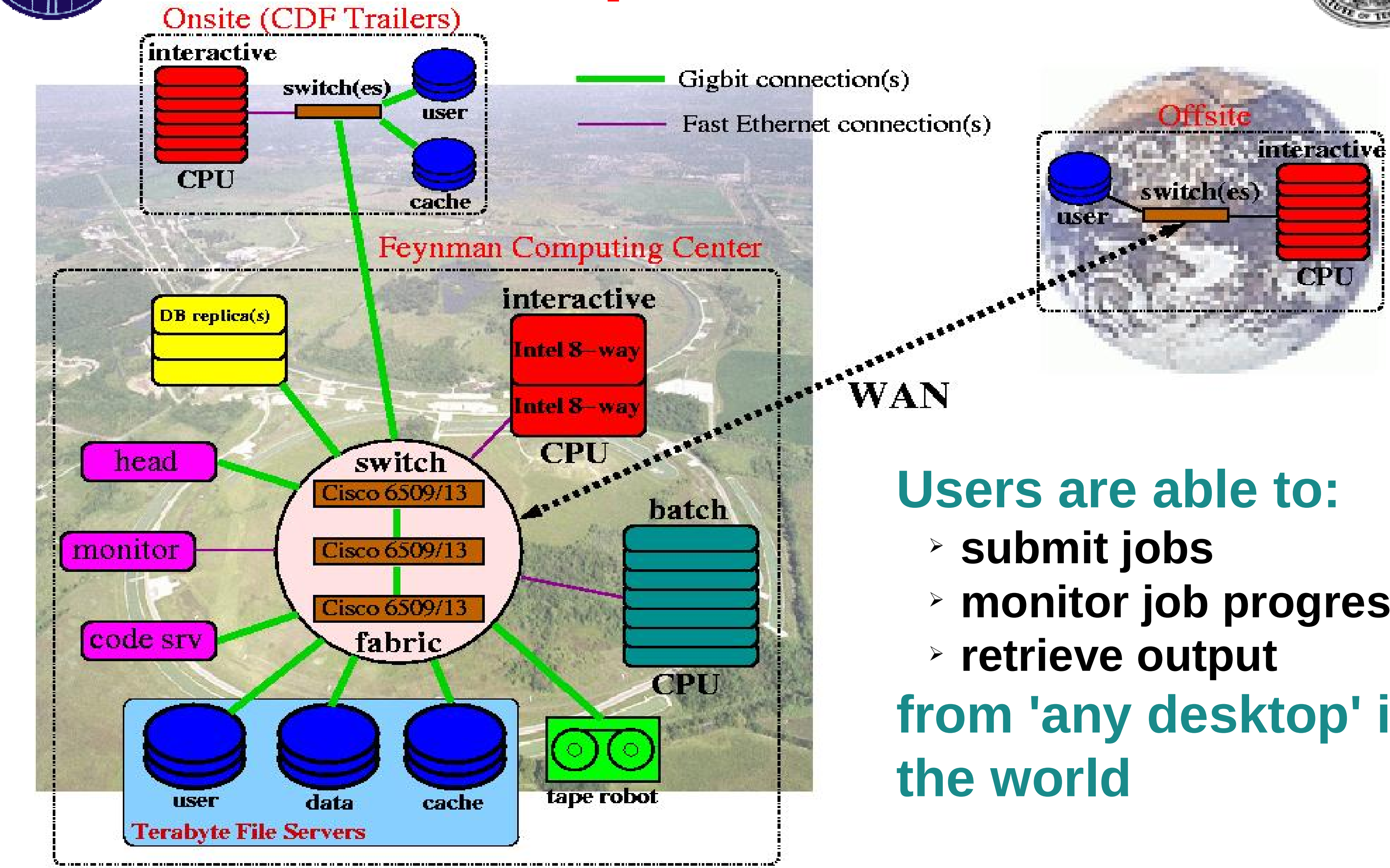
- Front-end disk cache for Enstore (read and write disk pools)
- Currently in β testing → working toward production use in CDF

SAM → framework for global data handling/distribution

- Jointly developed by FNAL Computing division and D0
- Works with Enstore and CDF analysis software framework
- Currently under evaluation for use in CDF data distribution

→ see Igor Terekhov's talk

CAF Implementation

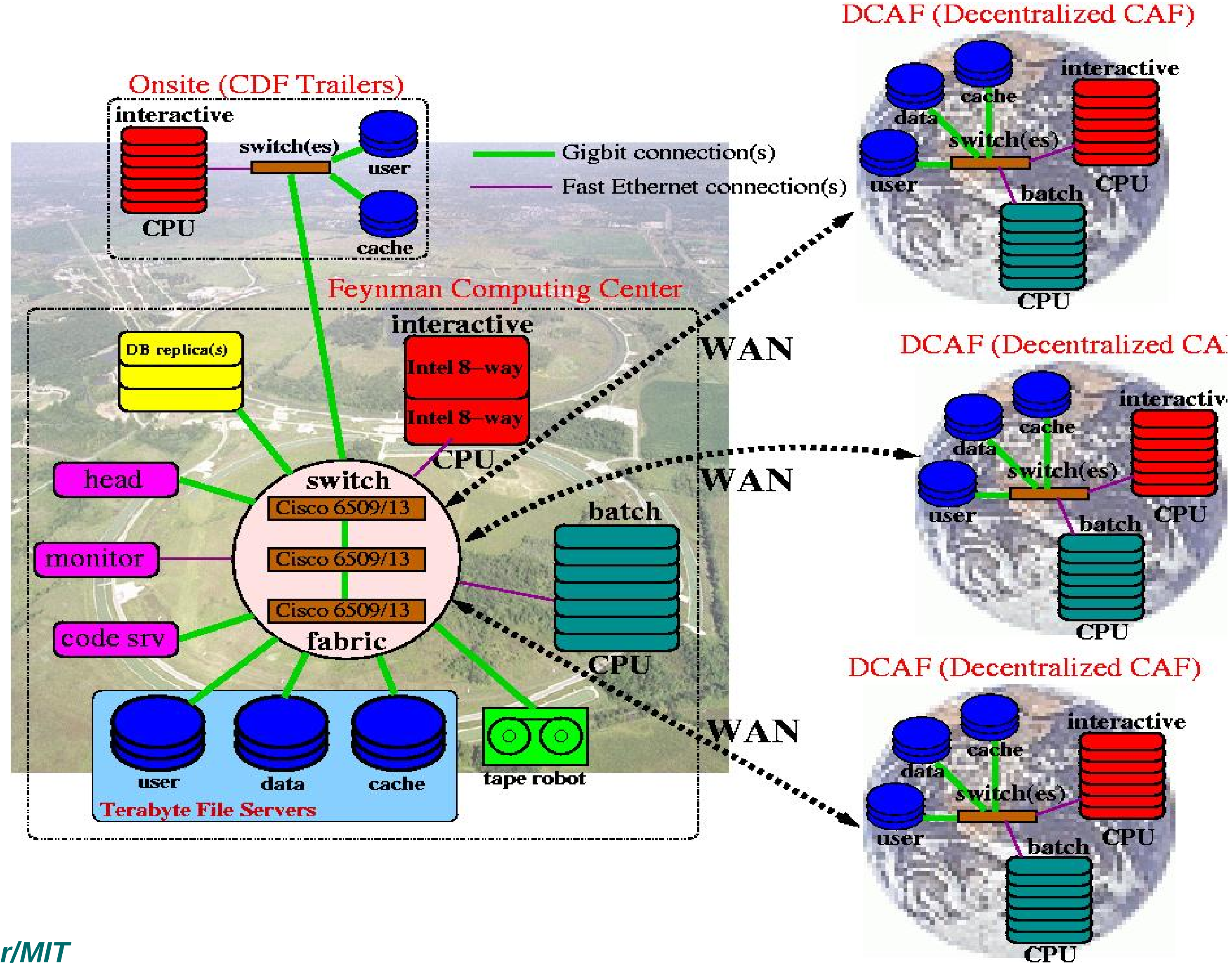


Users are able to:

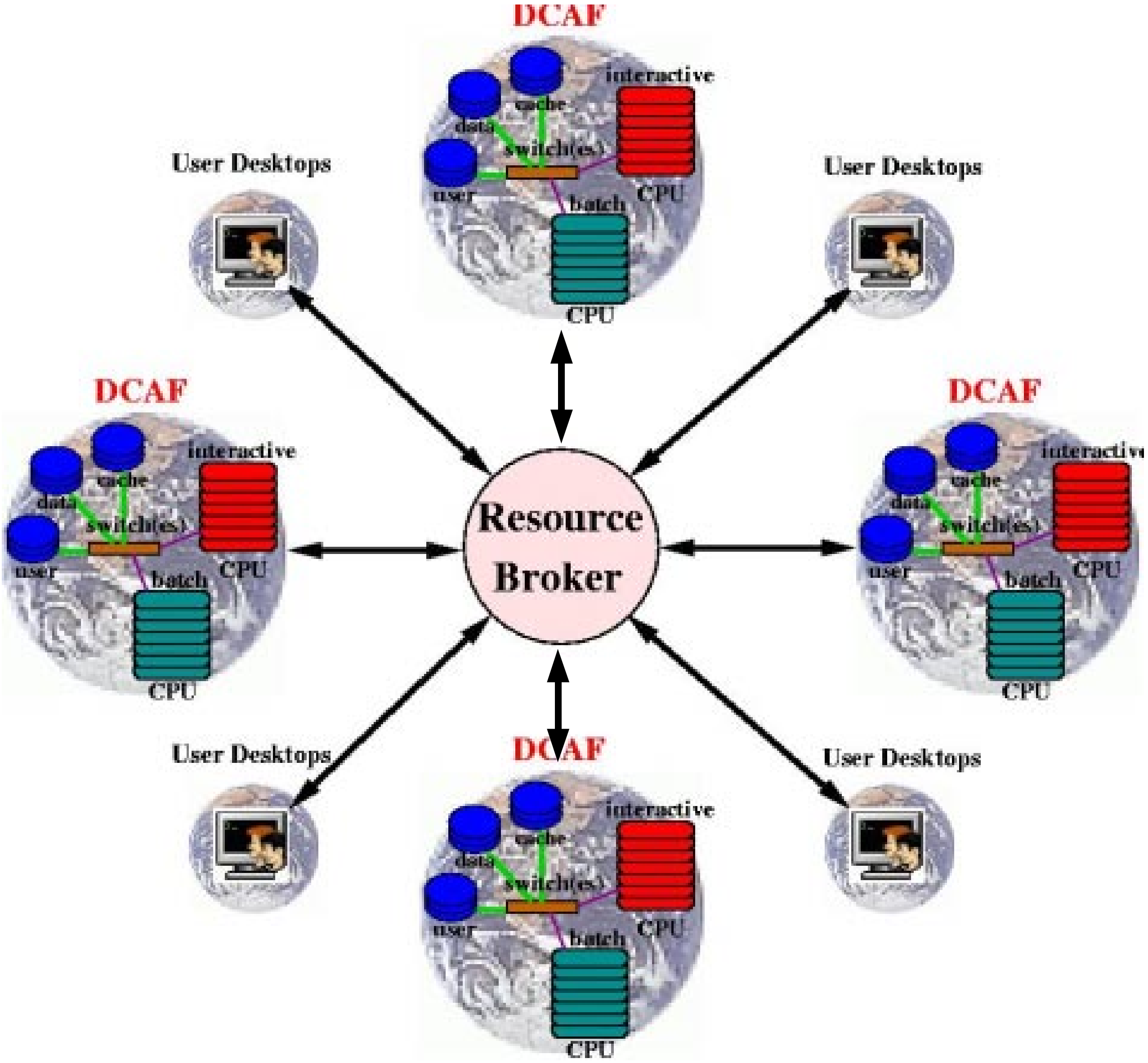
- submit jobs
- monitor job progress
- retrieve output

from 'any desktop' in the world

Toward the Grid



Peer-to-Farm Paradigm



Brokering scheme

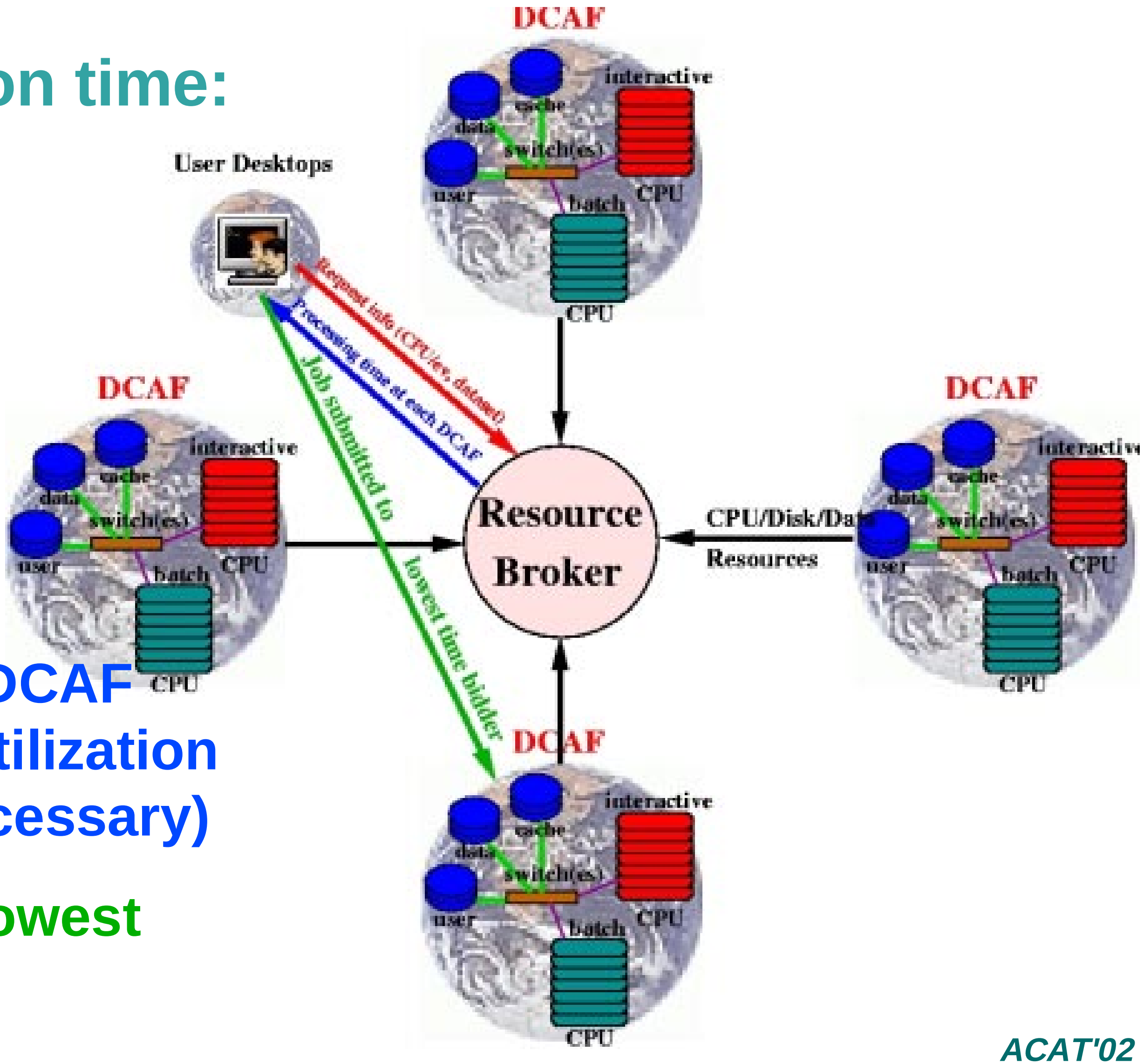
Minimize job execution time:

DCAFs update broker
- CPU/disk utilization
- Local data

User generates request
- CPU time/event
- Metadata ID (dataset)

Execution time on each DCAF
- CPU+I/O resources+utilization
- Data movement (if necessary)

Job goes to DCAF with lowest 'bid'



Summary/Conclusions

Distributed Peer-to-Farm Computing Model

Production system under heavy use:

- Single farm at FNAL
- Many peers all over the world
    - 100+ total users
    - 100+ simultaneous jobs
    - Regularly up to 800 jobs per user queued

Future development:

- Extend data handling
- Multi-farm brokering
- Scale system by O(10)